Mallows' C_p

Suppose that there are k predictors, $x_1, x_2, \ldots, x_k$, and that $E(Y|x_1, \ldots, x_k)$ is a linear fn of some subset of $x_1, \ldots, x_k$. For example, we could have $k = 9$, and

$$E(Y|x_1, x_2, \ldots, x_9) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4.$$

(Here the 9 predictors can be 9 different attributes, or we can have something like $x_1 = u$, $x_2 = v$, $x_3 = w$, $x_4 = u^2$, $x_5 = v^2$, $x_6 = w^2$, $x_7 = uv$, $x_8 = uw$, and $x_9 = vw$.) If we consider a specific linear model based on $p-1$ of the k predictors, and including an intercept term, so that there are p coefficients to estimate, let SSR denote the sum of the squared residuals which results from fitting the model to data, and let $\hat{\sigma}^2$ denote a good estimate of the error term variance (which is not necessarily SSR/(n-p)), the value of *Mallows'* C_p for this model is

$$C_p = \frac{SSR}{\hat{\sigma}^2} + 2p - n.$$

(Note: If the assumption that a linear model based on some subset of the k predictors is the correct model (but not nec. the best model to use in practice) is a correct assumption, then we could use $SSR_{full}/(n-k-1)$ far $\hat{\sigma}^2$, where SSR_{full} is the sum of the squared residuals resulting from fitting a linear model which uses all of the predictors, provided that the sample size is not too small.)

There seems to be quite a bit of confusion pertaining to the use of C_p. My guess is that this is because C_p can be used in two different ways! It can be used to assess biasedness in regression models, and it can also be used to select the model, from among a collection of candidates, which will hopefully minimize the mean squared error (prediction errors and/or error in estimating $E(Y \mid \vec{x})$). Some books and people seem to strongly focus on the bias issue. But I think the emphasis should be on the overall situation with error — the mean squared error takes into account error due to both bias and variance.

If an unbiased model is considered, we have that

$$E\left(\frac{SSR}{n-p}\right) = \sigma^2,$$

where σ^2 is the actual error term variance.

It follows that we have that

$$E(SSR) = (n-p)\sigma^2,$$

and so we ***might*** also have that

$$E(C_p) = E\left(\frac{SSR}{\hat{\sigma}^2} + 2p - n\right)$$

$$\approx (n-p) + 2p - n$$

$$= p.$$

So, if we have an unbiased model we expect to obtain a value of C_p which is close to p. A practice of some is that when considering a collection of models fit from the same data, focus should be on those having C_p values close to p, and of those no bias / low bias models, the one chosen to use should be the one having the smallest value of p. Favoring small p is due to the fact that unnecessary terms contribute to variance, and thus to error. But should the

focus only be on unbiased models?!?!!
(It's possible that a biased model can be better than the best unbiased model.)

Before moving on to the other use for C_p, let's consider again the situation described on p.1. The "correct" model is

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4,$$

and it should yield a value of C_p close to 4. The model

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

is also unbiased, because the true value of β_3 is 0, and so it should yield a value of C_p close to 5. The model

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2$$

is biased. It should yield a value of C_p larger than $p = 3$, because the biasedness gives us

$$E\left(\frac{SSR}{n-p}\right) > \sigma^2.$$

But if the biasedness is not too great, and the value of C_p is less than 4, the indication is that this biased model will be better than the "correct" model with regard to overall errors.

Here is another way of looking at C_p. Letting μ_i denote $E(Y_i \mid \vec{x}_i)$, and letting $\hat{\mu}_i$ be the OLS estimator of μ_i, C_p is an estimate of

$$\frac{\sum_{i=1}^n MSE(\hat{\mu}_i)}{\sigma^2}.$$

To me, this strongly suggests that one should select the model which results in the smallest value of C_p!!! (As I've said many times in STAT 554 and STAT 652, I think some put way too much emphasis on unbiasedness. Minimizing overall error should be the main thing, and one should take advantage of the bias-variance tradeoff in the best possible way.)

It should be noted that the C_p-like measure given on p. 203 of HTF is not the usual C_p:

$$C_{p\,(HTF)} = \frac{SSR}{n} + \frac{2p}{n}\hat{\sigma}^2 = \frac{\hat{\sigma}^2}{n} C_{p\,(usual)} + \hat{\sigma}^2.$$

$C_{p\,(HTF)}$ is an estimate of the average mean squared prediction error, based on a model having p terms, of n new observations taken at the points $\vec{x}_1, \vec{x}_2, \ldots, \vec{x}_{n-1}$, and $\vec{x}_n$.

The derivation of $C_{p\,(HTF)}$ is "sketched" in Sec. 7.4 and Sec. 7.5 of HTF, and doing Exercises 7.4 and 7.5 of HTF should add to one's understanding. C_p can be derived similarly.